Гребневая регрессия как байесовский метод

Ю. А. Калнишкан
yura@cs.rhul.ac.uk

Department of Computer Science
and Computer Learning Research Centre
Royal Holloway, University of London

2009

План доклада

- в докладе будет рассказано о
  — гребневой регрессии;
  — байесовской статистике (вкратце);
  — одном новом тождестве для гребневой регрессии и его следствиях;
  — связи с соединяющим алгоритмом Вовка
- Литература:
  — C. Bishop. Pattern recognition and machine learning. Springer, New York, 2006.
  — F. Zhdanov and V. Vovk, Competing with Gaussian linear experts, arXiv:0910.4683v1

Содержание

1. Гребневая регрессия

2. Кто такие байесовцы
(и почему они воюют против фриквентистов)

3. Регрессия с байесовской точки зрения

4. Тождество для секвенциальной постановки

5. То же с точки зрения соединяющего алгоритма Вовка (AA)

Постановка задачи

- пусть нам дана *обучающая выборка* (training set) $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$
  — $x_i \in \mathbb{R}^n$ *сигналы*
  — $y_i \in \mathbb{R}$ *метки* (labels) или *исходы* (outcomes)
- *задача регрессии*: нужно установить зависимость между исходами и сигналами, чтобы определять исходы для новых сигналов
- замечание: если y_i берутся из дискретного множества, то говорят о задаче классификации

Метод наименьших квадратов (1)

- ограничимся линейными зависимостями вида $y = \theta' x$, где $\theta \in \mathbb{R}^n$
  — будем называть θ *регрессорами*
  — нелинейный случай может быть сведён к линейному при помощи *кёрнельного трюка* (kernel trick)
- *метод наименьших квадратов (МНК)* (least squares): выбираем регрессор θ, минимизирующий суммарное квадратичное уклонение

$$\sum_{t=1}^{T} (\theta' x_t - y_t)^2$$

Метод наименьших квадратов (2)

- определим $T \times n$-матрицу $X = (x_1 x_2 \ldots x_T)'$ (строки X это векторы x_i')
- поиск θ по МНК

$$\sum_{t=1}^{T} (\theta' x_t - y_t)^2 = \|X\theta - Y'\|^2 \to \min_{\theta}$$

состоит в проецировании $Y = (y_1, y_2, \ldots, y_T)$ на плоскость, натянутую на столбцы X
— точка проекции единственна
— но если столбцы X линейно зависимы ($\operatorname{rank} X < n$), то точка проекции раскладывается по ним неоднозначно и θ не единственна
— если обучающая выборка мала ($T < n$), МНК не определяет регрессор однозначно

Гребневая регрессия

- добавим слагаемое $a\|\theta\|^2$, $a > 0$:

$$a\|\theta\|^2 + \sum_{t=1}^{T} (\theta' x_t - y_t)^2 \to \min_{\theta}$$

- можно сказать, что мы добавили n фиктивных примеров $(\sqrt{a} e_i, 0)$, где e_i – орты
  — вместо T обучающих примеров мы получаем $T + n$, причём n из них заведомо независимы, так что регрессор теперь определён однозначно
- нахождение θ по этому соотношению называется *гребневой регрессией* (ridge regression)
  — МНК можно рассматривать как предельный случай гребневой регрессии при стремлении a к 0

Обоснование

- Каков интуитивный смысл коэффициента a? Зачем его добавлять?
  — имеется две основных причины
- член $\|\theta\|^2$ можно считать сложностью гипотезы θ, а $\sum_{t=1}^{T}(\theta' x_t - y_t)^2$ точностью, с которой она описывает данные (goodness-of-fit)
- гребневая регрессия ищет баланс между точностью описания и сложностью
  — поэтому её можно рассматривать как MDL (minimum description length) метод
- вторая причина носит вычислительный характер и будет очевидна из дальнейшего

Нахождение минимума (1)

- минимум
$$L_{\mathrm{RR}}(\theta) = a\|\theta\|^2 + \|X\theta - Y'\|^2$$
ищем дифференцированием по θ
— когда θ уходит на бесконечность, $L_{\mathrm{RR}}(\theta) \to +\infty$, так что достаточно рассмотреть нули производной
- производная равна
$$\frac{d}{d\theta} L_{\mathrm{RR}}(\theta) = 2a\theta - 2X'Y + 2X'X\theta$$

Нахождение минимума (2)

- производная зануляется в единственной точке
$$\theta_{\mathrm{RR}} = (aI + X'X)^{-1}X'Y$$
— $n \times n$-матрица $X'X$ является симметричной неотрицательно определённой, но может вырождаться (напр., при $T < n$)
— добавляя слагаемое aI (при $a > 0$) мы делаем матрицу положительно определённой и невырожденной
- увеличивая a мы удаляемся от сингулярной матрицы, делая задачу более простой вычислительно
  — вот обещанная вторая причина введения $a > 0$

Дуальная форма

- предсказание гребневой регрессии на сигнале x можно записать в форме
$$\theta'_{\mathrm{RR}} x = Y'(aI + K)^{-1}k(x),$$
где
— $Y' = (y_1, y_2, \ldots, y_T)$
— $K = XX' = (\langle x_i, x_j \rangle)_{i,j=1}^{T}$: $T \times T$-матрица Грама векторов $x_1, x_2, \ldots, x_T$
— $k(x) = Xx = (\langle x_i, x \rangle)_{i=1}^{T}$
- это представление позволяет рассматривать нелинейную регрессию...

1. Гребневая регрессия

2. Кто такие байесовцы
(и почему они воюют против фриквентистов)

3. Регрессия с байесовской точки зрения

4. Тождество для секвенциальной постановки

5. То же с точки зрения соединяющего алгоритма Вовка (AA)

Оценка параметра

- пусть дана выборка $\mathcal{D} = (x_1, x_2, \ldots, x_T)$, $x_i \in \{0, 1\}$, полученная подбрасыванием кривой монеты
- требуется оценить вероятность выпадения орла $\mu = \Pr\{x = 1\}$
  — оценка вероятности успеха в схеме Бернулли

Оценка наибольшего правдоподобия

- традиционный (*фриквентистский*) подход:
  — давайте оптимизируем какой-нибудь критерий по μ
  — например, правдоподобие

$$\Pr(\mathcal{D} \mid \mu) = \mu^{\sharp_1 \mathcal{D}} (1 - \mu)^{\sharp_0 \mathcal{D}} \to \max_{\mu}$$

  — обозначение $\mid \mu$ здесь носит формальный характер
- взяв логарифм и приравняв к нулю производную, получаем оценку

$$\hat{\mu} = \frac{\sharp_1 \mathcal{D}}{\sharp \mathcal{D}}$$

т.е. μ оценивается как доля орлов

Избыточное обучение

- предположим, мы кидали монетку три раза и получили трёх орлов: $\mathcal{D} = \{1, 1, 1\}$
- оценка наибольшего правдоподобия даёт вероятность исхода

$$\hat{\mu} = \frac{3}{3} = 1$$

- это слишком сильное заключение
  — так называемое *избыточное обучение* (переобучение, overfitting)
- традиционная статистика говорит нам:
  — подобное случается редко; вероятность получить плохую выборку $\mathcal{D}$ при $\mu \neq 1$ мала
  — но нам приходится иметь дело с конкретной выборкой, а не со случайной...

Байесовский подход

- будем считать μ случайной величиной с *априорным распределением* (prior) с плотностью $p(\mu)$
- зная $p(\mu)$ и $\Pr(\mathcal{D} \mid \mu)$, можем подсчитать плотность *апостериорного распределения* (posterior)

$$p(\mu \mid \mathcal{D}) = \frac{\Pr(\mathcal{D} \mid \mu) p(\mu)}{\Pr(\mathcal{D})}$$

- знаменатель $\Pr(\mathcal{D})$ это нормализационный коэффициент — часто удобно написать

$$p(\mu \mid \mathcal{D}) \sim \Pr(\mathcal{D} \mid \mu) p(\mu)$$

и подобрать его позднее из условия

$$\int_{-\infty}^{+\infty} p(\mu \mid \mathcal{D}) d\mu = 1$$

Как выбрать априорное распределение?

- априорное распределение должно содержать в себе исходные представления о μ
  — закон жизни
  — тайну веков
- правдоподобие пропорционально $\Pr(\mathcal{D} \mid \mu) \sim \mu^x (1-\mu)^y$
- если априорное распределение имеет такую же форму, то апостериорное тоже будет такое
  — тип распределения будет сохраняться при увеличении выборки
  — в таком случае говорят о *сопряжённых распределениях*

Бета-распределение

- бета-распределение с параметрами $a, b > 0$ имеет плотность

$$p(\mu | a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1},$$

при $\mu \in [0, 1]$

- матожидание $\mathbf{E}\,\mu = \frac{a}{a+b}$
- мы выбрали это распределение из соображений удобства...
  — байесовцы так часто делают

Байесовское оценивание

- пусть априорное распределение это бета-распределение с параметрами a и b
- тогда апостериорное распределение это бета-распределение с параметрами $a + \sharp_1 \mathcal{D}$ и $b + \sharp_0 \mathcal{D}$

$$p(\mu \mid \mathcal{D}) = \frac{\Gamma(a+b+\sharp\mathcal{D})}{\Gamma(a+\sharp_1\mathcal{D})\Gamma(b+\sharp_0\mathcal{D})} \mu^{a+\sharp_1\mathcal{D}-1} (1-\mu)^{b+\sharp_0\mathcal{D}-1}$$

- как выбрать точечную оценку?
- среднеквадратичные потери $\mathbf{E}(\tilde{\mu} - \mu)^2$ достигают минимума при $\tilde{\mu} = \mathbf{E}\,\mu$
- итак, берём

$$\tilde{\mu} = \frac{a + \sharp_1 \mathcal{D}}{a + b + \sharp\mathcal{D}}$$

Связь с оценкой наибольшего правдоподобия

$$\frac{a+\sharp_1\mathcal{D}}{a+b+\sharp\mathcal{D}} = \frac{\sharp_1\mathcal{D}}{\sharp\mathcal{D}}\cdot\frac{\sharp\mathcal{D}}{a+b+\sharp\mathcal{D}} + \frac{a}{a+b}\cdot\frac{a+b}{a+b+\sharp\mathcal{D}}$$
$$\tilde{\mu} = \hat{\mu}(1-\lambda) + \lambda\mu_0,$$

где $\lambda \in [0,1]$ и $\lambda \to 0$ при $\sharp\mathcal{D} \to \infty$

- байесовская оценка это смесь оценки наибольшего правдоподобия с априорной
- при увеличении размера выборки байесовская оценка стремится к оценке наибольшего правдоподобия
- байесовский подход даёт более консервативные оценки и позволяет бороться с излишним обучением

Модель

- примем следующую модель:

$$y_t = \theta' x_t + \varepsilon_t,$$

где ε_t независимы с одинаковым распределением $\mathcal{N}(0,\sigma^2)$

- для y_t получаем плотность

$$p(y_t \mid x_t, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\theta' x_t - y_t)^2}$$

- совместная плотность (в силу независимости)

$$p(y_1, y_2, \ldots, y_T \mid x_1, x_2, \ldots, x_T, \theta) = \frac{1}{(2\pi\sigma^2)^{T/2}} e^{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(\theta' x_t - y_t)^2}$$

— впредь $x_1, x_2, \ldots, x_T$ мы писать не будем для краткости

Оценка наибольшего правдоподобия

- максимизация правдоподобия

$$p(y_1, y_2, \ldots, y_T \mid \theta) \to \max_\theta$$

равносильна минимизации

$$\sum_{t=1}^{T}(\theta' x_t - y_t)^2 \to \min_\theta$$

- итак, оценка наибольшего правдоподобия это метод наименьших квадратов

Сопряжённое распределение

- правдоподобие пропорционально $e^{-Q(\theta)}$, где
  — $Q(\theta) = \theta' A\theta + B\theta + C$ квадратичная функция
  — $A = \sum x_t x_t'$ симметричная неотрицательно определённая матрица
- возьмём априорное распределение вида

$$p(\theta) \sim e^{-R(\theta)},$$

  — где квадратичная часть $R(\theta)$ положительно определена (а то интеграл не сойдётся)
  – такая форма распределения будет сохранятся при переходе от априорной к апостериорной плотности

Апостериорное распределение

- возьмём априорную плотность

$$p(\theta) = \left(\frac{a}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{a}{2\sigma^2}\|\theta\|^2} = \left(\frac{a}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2}\theta'(\frac{a}{\sigma^2}I)\theta}$$

- получаем апостериорное распределение

$$\begin{aligned} p(\theta \mid y_1, y_2 \ldots, y_T) &\sim p(\theta) p(y_1, y_2, \ldots, y_T \mid \theta) \\ &\sim e^{-\frac{1}{2\sigma^2}\left(\sum_{t=1}^{T}(\theta' x_t - y_t)^2 + a\|\theta\|^2\right)} \end{aligned}$$

- это, очевидно, нормальное распределение

Нормальное распределение общего вида

- плотность общего n-мерного нормального распределения

$$p(\theta) = \frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{\sqrt{\det \Sigma}} e^{-\frac{1}{2}(\theta-\mu)'\Sigma^{-1}(\theta-\mu)}$$

- здесь
  — $\mathbf{E}\,\theta = \mu$ – среднее; оно совпадает с модой, т.е. точкой максимума плотности (можно проверить дифференцированием)
  — $\mathbf{var}\,\theta = \Sigma$ – матрица ковариаций

Параметры апостериорного распределения

- для апостериорного распределения регрессора θ получаем
  — среднее равно моде и получается минимизацией

$$\sum_{t=1}^{T}(\theta' x_t - y_t)^2 + a\|\theta\|^2 \to \min_{\theta}$$

  — матрицу ковариаций можно подсчитать, выделив квадратичную часть

$$\Sigma^{-1} = \frac{1}{\sigma^2}\left(\sum_{t=1}^{T} x_t x_t' + aI\right)$$

- итак, апостериорное распределение это

$$\mathcal{N}\left(\theta_{\mathrm{RR}}, \sigma^2\left(\sum_{t=1}^{T} x_t x_t' + aI\right)^{-1}\right)$$

Секвенциальная постановка

- рассмотрим *секвенциальный* (последовательный, on-line) протокол:

  FOR $t = 1, 2, \ldots$
  (1) наблюдаем сигнал $x_t \in \mathbb{R}^n$
  (2) выдаём предсказание $\gamma_t \in \Gamma = \mathbb{R}$
  (3) наблюдаем исход $y_t \in \Omega = \mathbb{R}$
  END FOR

- мы одновременно изучаем зависимость y от x и пытаемся её предсказать
- предыдущая постановка, при которой все пары (x_t, y_t) даны сразу, называется *пакетной* (batch)

Гребневая регрессия в секвенциальной постановке

- гребневую регрессию, как и любой другой пакетный метод, можно применять в секвенциальной постановке
- на шаге t мы
  — составляем обучающую выборку из известных пар $(x_1, y_1), (x_2, y_2), \ldots, (x_{t-1}, y_{t-1})$
  — находим θ_{t-1}
  — предсказываем $r_t = \theta'_{t-1} x_t$
- накопленные квадратичные потери $\sum_{t=1}^{T} (r_t - y_t)^T$ вообще говоря превосходят минимум

$$\min_{\theta} \sum_{t=1}^{T} (\theta' x_t - y_t)^2$$

Байесовский метод в секвенциальной постановке

- возьмём априорное распределение $p_0(\theta)$
- после $t-1$ шага имеем апостериорное распределение $p_{t-1}(\theta) = p(\theta \mid y_1, y_2, \ldots, y_{t-1})$
  — используем его как-нибудь для получения предсказания γ_t
- получив y_t и приняв $p_{t-1}(\theta)$ за априорное распределение, построим апостериорное распределение $p_t(\theta) = p(\theta \mid y_1, y_2, \ldots, y_{t-1}, y_t)$

Применение байесовского метода: θ

- возьмём априорное распределение

$$p(\theta) = \left(\frac{a}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{a}{2\sigma^2}\|\theta\|^2}$$

- правдоподобие на шаге t даётся плотностью

$$p(y_t \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\theta' x_t - y_t)^2}$$

- итого для апостериорного распределения θ имеем как и раньше

$$\begin{aligned} p_{T-1}(\theta) &\sim p(\theta)p(y_1 \mid \theta)p(y_2 \mid \theta)\dots p(y_{T-1} \mid \theta) \\ &\sim e^{-\frac{1}{2\sigma^2}\left(\sum_{t=1}^{T-1}(\theta' x_t - y_t)^2 + a\|\theta\|^2\right)} \end{aligned}$$

Применение байесовского метода: предсказание

- среднее θ_{T-1} – такое же, как при пакетной постановке
- для $y_T = \theta' x_T + \varepsilon_T$ получаем распределение $\mathcal{N}(r_T, \sigma_T^2)$, где
  — $r_T = \theta'_{T-1} x_T$
  — $\sigma_T^2 = \sigma^2 x_T' \left(\sum_{i=1}^{T-1} x_i x_i' + aI\right)^{-1} x_T + \sigma^2 =$
  $\sigma^2(x_T' A_{T-1}^{-1} x_T + 1)$, где $A_t = \sum_{i=1}^{t} x_i x_i' + aI$
- можем выдать в качестве предсказания среднее r_T
- гребневая регрессия в секвенциальной постановке снова оказывается байесовским методом

Совместное распределение

- подсчитаем полное совместное распределение $(y_1, y_2, \dots, y_T)$
  — можно считать, что мы смотрим «из будущего» и последовательность сигналов x_t нам известна
- подсчитаем плотность двумя способами

Цепочка условных вероятностей

- представим вероятность так:

$$\begin{aligned} &p(y_1, y_2, \dots, y_T) = \\ &p(y_T \mid y_1, y_2, \dots, y_{T-1})p(y_{T-1} \mid y_1, y_2, \dots, y_{T-2}) \dots p(y_2 \mid y_1)p(y_1) \end{aligned}$$

- значения $y_1, y_2, \dots, y_{T-1}$ определяют распределение для θ и, соответственно, для y_T, так что

$$p(y_T \mid y_1, y_2, \dots, y_{T-1}) = \frac{1}{\sqrt{2\pi\sigma_T^2}} e^{-\frac{1}{2\sigma_T^2}(r_T - y_T)^2}$$

- итого

$$p(y_1, y_2, \dots, y_T) = \frac{1}{(2\pi)^{T/2}\sigma_1\sigma_2\dots\sigma_T} e^{-\sum_{t=1}^{T}\frac{1}{2\sigma_t^2}(r_t - y_t)^2}$$

Маргинализация (1)

- представим плотность как интеграл от более общего распределения

$$p(y_1, y_2, \ldots, y_T) = \int_{\mathbb{R}^n} p(y_1, y_2, \ldots, y_T, \theta) d\theta$$

- имеем

$$p(y_1, y_2, \ldots, y_T, \theta) = p(y_1, y_2, \ldots, y_T \mid \theta) p(\theta)$$

Маргинализация (2)

- при данном θ все y_t независимы, т.е.

$$\begin{aligned} p(y_1, y_2, \ldots, y_T \mid \theta) &= p(y_1 \mid \theta) p(y_2 \mid \theta) \ldots (y_T \mid \theta) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_1 - \theta' x_1)^2} \ldots \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_T - \theta' x_T)^2} \\ &= \frac{1}{(2\pi)^{T/2}\sigma^T} e^{-\frac{1}{2\sigma^2}\sum_{t=1}^T (y_t - \theta' x_t)^2} \end{aligned}$$

$$p(y_1, y_2, \ldots, y_T, \theta) = \frac{1}{(2\pi)^{T/2}\sigma^T} \cdot \frac{a^{n/2}}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\left(\sum_{t=1}^T (y_t - \theta' x_t)^2 + a\|\theta\|^2\right)}$$

- необходимо подсчитать интеграл по $d\theta$

Интеграл

- пусть $Q(x) = x'Ax + Bx + C$, где $x \in \mathbb{R}^n$ и A – симметричная положительно определённая матрица — тогда

$$\int_{\mathbb{R}^n} e^{-Q(x)} dx = e^{-Q_0} \frac{\pi^{n/2}}{\sqrt{\det A}}$$

  где $Q_0 = \min_x Q(x)$
- можно получить сведением к результатам о нормальном распределении

Сравнение

- приравняем минус логарифмы вероятности

$$\begin{aligned} &\sum_{t=1}^T \frac{1}{2\sigma_t^2}(r_t - y_t)^2 + \sum_{t=1}^T \ln \sigma_t = \\ &\qquad \frac{1}{2\sigma^2} \min_\theta \left(\sum_{t=1}^T (y_t - \theta' x_t) + a\|\theta\|^2 \right) - \\ &-\frac{n}{2}\ln\pi + \frac{1}{2}\ln\det\left(\frac{1}{2\sigma^2}\left(\sum_{t=1}^T x_t x_t' + aI\right)\right) + \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln a + n\ln\sigma \end{aligned}$$

- слева

$$\sum_{t=1}^T \ln \sigma_t = \frac{1}{2}\sum_{t=1}^T \ln \sigma_t^2 = T\ln\sigma + \frac{1}{2}\sum_{t=1}^T \ln(x_t' A_{t-1} x_t + 1)$$

Упрощение детерминанта (1)

- сначала вынесем множитель

$$\frac{1}{2}\ln\det\left(\frac{1}{2\sigma^2}\left(\sum_{t=1}^{T}x_t x_t' + aI\right)\right) = \frac{1}{2}\ln\det\left(\sum_{t=1}^{T}\frac{1}{a}x_t x_t' + I\right) + \frac{1}{2}\ln\left(\frac{a}{2\sigma^2}\right)^n$$

— второе слагаемое сократится с мелкими членами в формуле, а

$$\det\left(\sum_{t=1}^{T}\frac{1}{a}x_t x_t' + I\right) = \frac{1}{a^n}\det A_T$$

Упрощение детерминанта (2)

- применим тождество Сильвестра $\det(I + AB) = \det(I + BA)$

$$\det A_T = \det\left(\sum_{t=1}^{T}x_t x_t' + aI\right) = \det(x_T x_T' + A_{T-1}) =$$
$$\det[(x_T x_T' A_{T-1}^{-1} + I)A_{T-1}] = \det A_{T-1}\det(x_T x_T' A_{T-1}^{-1} + I) =$$
$$\det(aI_n)\prod_{t=1}^{T}\det(x_t x_t' A_{t-1}^{-1} + I) = a^n\prod_{t=1}^{T}(x_t' A_{t-1}^{-1} x_t + 1)$$

- всё сокращается!

Тождество

- получаем

$$\sum_{t=1}^{T}\frac{(r_t - y_t)^2}{1 + x_t A_{t-1}^{-1} x_t} = \min_{\theta}\left(\sum_{t=1}^{T}(y_t - \theta' x_t)^2 + a\|\theta\|^2\right)$$

где $A_t = \sum_{i=1}^{t} x_i x_i' + aI$
— σ исчезла из равенства, так что вероятностные предположения можно отбросить
— слева стоят почти потери гребневой регрессии в секвенциальной постановке
— справа величина «потери + сложность» для ретроспективно лучшего регрессора

- нас больше интересуют потери $\sum_{t=1}^{T}(r_t - y_t)^2$
— величина $x_t A_{t-1}^{-1} x_t > 0$ не очень значительна

Следствие (1)

- оценим

$$\frac{1}{1+b} = 1 - \frac{b}{1+b} \geq 1 - \ln(1+b)$$

(здесь $b \geq -1$)
— имеем $\sum_{t=1}^{T}\ln\left(1 + x_t A_{t-1}^{-1} x_t\right) = \ln\det(A_T/a)$

- ограничим интервал
— будем считать, что $y_i \in [-Y, Y]$, и этот интервал нам заранее известен
— будем «подрезать» предсказания гребневой регрессии r_t так, чтобы они попали в интервал
— для подрезанных предсказаний $\bar{r}_t$ имеем $(\bar{r}_t - y_t)^2 \leq 4Y^2$ и $(\bar{r}_t - y_t)^2 \leq (r_t - y_t)^2$

Следствие (2)

- получаем

$$\sum_{t=1}^{T}(\bar{r}_t - y_t)^2 \le \min_\theta \left(\sum_{t=1}^{T}(y_t - \theta' x_t)^2 + a\|\theta\|^2 \right) + 4Y^2 \ln\det\left(I + \frac{1}{a}\sum_{t=1}^{T} x_t' x_t \right)$$

- детерминант симметричной положительно определённой матрицы не больше произведения диагональных элементов
- предположим, что $\|x_t\|_\infty \le X$, т.е. все координаты сигналов не превосходят X; тогда

$$\det\left(I + \frac{1}{a}\sum_{t=1}^{T} x_t' x_t \right) \le \left(1 + \frac{TX^2}{a} \right)^n$$

Следствие (3)

- получаем

$$\sum_{t=1}^{T}(\bar{r}_t - y_t)^2 \le \min_\theta \left(\sum_{t=1}^{T}(y_t - \theta' x_t)^2 + a\|\theta\|^2 \right) + 4Y^2 n \ln\left(1 + \frac{TX^2}{a} \right)$$
$$\le \min_\theta \left(\sum_{t=1}^{T}(y_t - \theta' x_t)^2 + a\|\theta\|^2 \right) + O(\ln T)$$

Оптимальность (1)

- есть пример, когда x_t одномерны (но неограничены) и
  — $\sum_{t=1}^{T}(\bar{r}_t - y_t)^2 = 4T + o(T)$
  — $\min_\theta \left(\sum_{t=1}^{T}(y_t - \theta x_t)^2 + a\theta^2 \right) \le T$
  — $\ln\det\left(I + \frac{1}{a}\sum_{t=1}^{T} x_t^2 \right) = 2T\ln 2 + O(1)$
  [Vovk, Competitive On-line Statistics, 2001]
- полученная оценка на потери гребневой регрессии довольно точна
  — но гребневая регрессия не оптимальна как метод секвенциального предсказания

Оптимальность (2)

- применяя соединяющий алгоритм Вовка, можно построить алгоритм с оценкой

$$\sum_{t=1}^{T}(\tilde{r}_t - y_t)^2 \le \min_\theta \left(\sum_{t=1}^{T}(y_t - \theta' x_t)^2 + a\|\theta\|^2 \right) + Y^2 \ln\det\left(I + \frac{1}{a}\sum_{t=1}^{T} x_t' x_t \right)$$

- существует стохастическая стратегия для природы в ситуации $\|x\|_\infty = 1$ и $Y = 1$, так что для любого алгоритма

$$\mathsf{E}\left(\sum_{t=1}^{T}(r_t - y_t)^2 - \min_\theta \left(\sum_{t=1}^{T}(y_t - \theta' x_t)^2 \right) \right) \ge (n - \varepsilon)\ln T + O(1)$$

1. Гребневая регрессия

2. Кто такие байесовцы
(и почему они воюют против фриквентистов)

3. Регрессия с байесовской точки зрения

4. Тождество для секвенциальной постановки

5. То же с точки зрения соединяющего алгоритма Вовка (AA)

Постановка задачи

- протокол для *предсказателя* $\mathfrak{M}$

  FOR $t = 1, 2, \ldots$
  (1) $\mathfrak{M}$ считывает сигнал x_t
  (2) $\mathfrak{M}$ считывает предсказания *экспертов* $\gamma_t^{(\theta)}$, $\theta \in \Theta$
  (3) $\mathfrak{M}$ выдаёт предсказание $\gamma_t \in \Gamma$
  (4) $\mathfrak{M}$ наблюдает исход $\omega_t \in \Omega$
  END FOR

- качество предсказаний меряется *функцией потерь* $\lambda(\omega, \gamma)$
- цель: соревноваться с лучшим экспертом по *накопленным потерям* $\mathrm{Loss}(T) = \sum_{t=1}^{T} \lambda(\omega_t, \gamma_t)$

Предсказание с помощью распределений

- пусть пространство исходов это $\Omega = \mathbb{R}$
- предсказания это непрерывные функции плотности

$$\Gamma = \left\{ q \in C(\mathbb{R}) \mid \int_{-\infty}^{+\infty} q(x)dx = 1 \right\}$$

- а потери это

$$\lambda(y, q) = -\ln q(y)$$

Соединяющий алгоритм Вовка

- соединяющий алгоритм Вовка (Aggregating Algorithm) позволяет смешивать предсказания экспертов, если игра достаточно «хорошая»
  — а тут она хорошая
- смешиваются потери в экспоненциальном пространстве

$$\alpha e^{-\eta(-\ln q_1)} + (1-\alpha)e^{-\eta(-\ln q_2)} = \alpha q_1^{\eta} + (1-\alpha)q_2^{\eta}$$

— для данной игры оптимальное значение $\eta = 1$

Вычисление веса

- экспертам назначаются веса, которые меняются так:

$$p_t^{(\theta)} = e^{-\eta\lambda(\omega_t,\gamma_t^{(\theta)})} p_{t-1}^{(\theta)} = q_t(y_t) p_{t-1}^{(\theta)}$$

— вес эксперта θ после шага t соответствует плотности $p(y_1, y_2, \ldots, y_t|\theta)p(\theta)$
— или доле денег в распоряжении эксперта θ

Выбор конкретных экспертов

- пусть эксперт $\theta \in \mathbb{R}^n$ выдаёт плотность нормального распределение

$$q_t(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\theta' x_t)^2}$$

- а начальное распределение весов на экспертах

$$p(\theta) = \left(\frac{a}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{a}{2\sigma^2}\|\theta\|^2}$$

— столько денег мы даём эксперту θ

Аналог цепочки условных вероятностей

- смешивание предсказаний экспертов в соответствии с соединяющим алгоритмом даст нам на шаге t распределение

$$\frac{1}{\sqrt{2\pi\sigma_t^2}} e^{-\frac{1}{2\sigma_t^2}(y-r_t)^2},$$

где r_t – предсказание гребневой регрессии

- величина

$$\frac{1}{(2\pi)^{T/2}\sigma_1\sigma_2\ldots\sigma_T} e^{-\sum_{t=1}^{T}\frac{1}{2\sigma_t^2}(r_t-y_t)^2}$$

теперь имеет смысл экспоненты от потерь предсказателя
— капитал предсказателя

Аналог маргинализации

- интеграл

$$\int_{\mathbb{R}^n} \frac{1}{(2\pi)^{T/2}\sigma^T} \cdot \frac{a^{n/2}}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\left(\sum_{t=1}^{T}(y_t-\theta' x_t)^2 + a\|\theta\|^2\right)} d\theta$$

имеет смысл смеси (экспоненциальных) потерь экспертов
— т.е. суммы капиталов экспертов

- эта игра линейна и капитал предсказателя равен сумме капиталов экспертов
- можем приравнять эти выражения

Заключение

- соединяющий алгоритм Вовка позволяет провести эквивалентное рассуждение на языке, не включающем байесовских предположений
- такой подход во многом более интуитивен — хотя на данный момент и известен гораздо меньшему количеству людей